Reply to Reviewer 1. Review is quoted in italics, with our responses interleaved.

The manuscript by An[n]an et al. is a useful addition to the literature, clearly laying out the issues of model uncertainty and the relatively sparse data available to compare to those models. The methodology used has been used before, by the authors, to reconstruct LGM conditions. Here they apply the same methodology to the mid Piacenzian (Pliocene).

-> Thank you for the comments.

The paper is a useful example of the methodology and poses questions that should be followed up by those looking at Pliocene and other deep-time climates. There are a number of minor issues, enumerated below, that detract from the overall message. If one is going to compare different proxy data sets, an attempt should be made to use as close to apples vs apples as one can get. A comparison of the PlioVAR and PRISM allkenone compilations, which use basically the same data, would be more informative if the same resolution was chosen. Instead of comparing the data from ± 10 kya windows around 3.205 Ma, a comparison was made using ±10 kya for one data set and ±15 kya for the other. This may not make much difference but could have been avoided.

-> These data sets have been made available to the modelling community as products for comparison with and validation of models. Our aim in selecting them was to investigate their implications for reconstruction of global temperature fields, rather than for the purposes of direct comparison. The windows are different because this is the nature of the data sets that have been provided to the modelling community (e.g. Haywood et al., 2020 uses the ±15 ky window of Foley and Dowsett for PlioMIP2 data-model comparison, whereas McClymont et al., 2020 focus on the ±10 ky window). Although the Foley and Dowsett database made available a 10K window which aligns with PlioVAR, the stratigraphic age controls on the sites used by PlioVAR were also reviewed and, in some cases, revised (outlined in McClymont et al., 2020) which also introduces differences between the two data sets. As the reviewer notes in their next comment, the impact of choosing the different data sets is minor, so we prefer not to have a detailed discussion about possible influences on differences which could detract from our main message.

Figure 4 shows anomaly maps for (a) PRISM4, (b) PlioVAR Mg/Ca and (c) PlioVAR ALL. It would be helpful to see a plot of PlioVAR Uk37 for comparison (I think this is Figure 1 (b)). Having it side by side as part of Figure 4 would make visual comparison of the different data sets more productive. The

differences between PlioVAR Uk37 and PRISM4 are minor, and both show marked differences compared to PlioVAR Mg/Ca. This isn't a surprise and is nicely documented quantitatively, but seeing adjacent images would help.

-> We will duplicate the plot from Fig 1 in Fig 4 if editorial staff allow.

The conclusion that the models may be underestimating polar amplification isn't much of a surprise to the community, but it is useful to document it as the authors have. Likewise, much is made of the mismatch between Mg/Ca and alkenone based SST estimates. This is nothing new, having been discussed in more detail in countless previous papers.

-> We agree with your comments about polar amplification.
As for the data mismatch, yes we agree this is not new but hope it is useful to emphasise the importance of this issue. The use of model fields to interpolate between sparse data points allows for a more comprehensive comparison across sites that are not co-located, when compared to a direct comparison of data points alone.

Thank you also for pointing out various mistakes in the references, which we will tidy up.

page 4 line 2: This appears to be a simple misunderstanding, but Bragg (2014) could not have used PRISM4 data since those data were not available prior to 2016 and SST estimates shown in Foley and Dowsett (referred to here as PRISM4), were not produced until 2019.

-> Yes, this will be corrected. Bragg used the PRISM2 and PRISM3 "time slabs" rather than the single interglacial KM5c (PRISM4).

Page 4, lines 18-19: Why use the PRISM4 community sourced verification data with a 30K window to compare to PLIOVAR's 20K window when PRISM also, in the same release, produced a version with a ±10K window?

-> In PlioMIP2 data-model comparisons (e.g. Haywood et al., 2020) which use Foley and Dowsett, 2019) the data set is referred to as the “PRISM4 SST data”. In the original definition of this data set (Figure 1, Dowsett et al., 2016) the PRISM4 time series is defined as encompassing Marine Isotope Stage M2 through to interglacial KM3 (3.190 to 3.220 Ma) which is the 30 kyr we have used in the manuscript.

Page 4, lines 28-29: The PlioVAR interval is slightly narrower only due to your choice of the 30 kyr window rather than the identical 20 kyr window provided in Foley and Dowsett (2019).

-> please see our reply to the previous question.

Page 5, lines 26-27: You should probably cite a couple of the many available references that previously documented differences between Mg/Ca and alkenone based SST estimates in Pliocene and Pleistocene sequences.

-> Agreed, we can add a line here to say “as observed in some of the original time series” and provide references to support the statement.

Page 5, lines 30-34: This is an interesting point and it would be helpful if it was addressed in this paper. Foley and Dowsett (2019) is a compilation of previously published alkenone data and it would be useful to know whether the sites not in common with PlioVAR are from a particular region, particular lab, etc.

We can provide a note to this effect in our revised manuscript. Broadly speaking, the sites which were not included in PlioVAR were the result of not meeting the PlioVAR time resolution constraints required for the data and/or the age model, and tend to be from studies where the focus was on low-resolution analysis of longer time series rather than any regional or lab specific bias.

Page 7, lines 33-35: As in one of the comments above, comparing alkenone and Mg/Ca based SST estimates is like apples and oranges. They are measuring different things and while both are calibrated to mean annual SST, the literature is ripe with examples of the two providing discordant estimates. On page 8 of this manuscript you indicate some possible reasons for Mg/Ca data being less reliable, the same reasons that have been stated by many authors in the past. Maybe move those up to page 7 and provide citations to earlier works?

Yes we can move some of the reasons up to an earlier part of the manuscript and provide some of the citations to earlier works. A challenge is that this is a complex subject which has also been reviewed and discussed extensively in other papers, including for the reasons given by the reviewer. We will ensure that relevant citations can direct readers to this issue. Our focus here was to see what impact including “all” or “selected” proxy data had on the model results rather than explaining proxy differences.